scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI

Year: 2023 | *Citations: 100* | *Authors: Haotian Cui, Chloe Wang, Hassaan Maan*

Abstract

Generative pre-trained models have achieved remarkable success in various domains such as natural language processing and computer vision. Specifically, the combination of large-scale diverse datasets and pre-trained transformers has emerged as a promising approach for developing foundation models. Drawing parallels between linguistic constructs and cellular biology — where texts comprise words, similarly, cells are defined by genes — our study probes the applicability of foundation models to advance cellular biology and genetics research. Utilizing the burgeoning single-cell sequencing data, we have pioneered the construction of a foundation model for single-cell biology, scGPT, which is based on generative pre-trained transformer across a repository of over 33 million cells. Our findings illustrate that scGPT, a generative pre-trained transformer, effectively distills critical biological insights concerning genes and cells. Through the further adaptation of transfer learning, scGPT can be optimized to achieve superior performance across diverse downstream applications. This includes tasks such as cell-type annotation, multi-batch integration, multi-omic integration, genetic perturbation prediction, and gene network inference. The scGPT codebase is publicly available at https://github.com/bowang-lab/scGPT.